Automated Detection of Fake Biomedical Papers: A Machine Learning Perspective

Ahmar K. Hussain[a], Bernhard A. Sabel[b], Marcus Thiel[c] and Andreas Nürnberger[d]
Otto von Guericke University Magdeburg, Germany
{*ahmar.hussain, marcus.thiel, andreas.nuernberger*}*@ovgu.de, bernhard.sabel@med.ovgu.de*

Keywords: Fake Papers, Classification, Meta Data Features, TF-IDF, Biomedicine, Large Language Models.

Abstract: In order to address the issue of fake papers in scientific literature, we propose a study focusing on the classification of fake papers based on certain features, by employing machine learning classifiers. A new dataset was collected, where the fake papers were acquired from the Retraction Watch database, while the non-fake papers were obtained from PubMed. The features extracted for classification included metadata, journal-related features as well and textual features from the respective abstracts, titles, and full texts of the papers. We used a variety of different models to generate features/word embeddings from the abstracts and texts of the papers, including TF-IDF and different variations of BERT trained on medical data. The study compared the results of different models and feature sets and revealed that the combination of metadata, journal data, and BioBERT embeddings achieved the best performance with an accuracy and recall of 86% and 83% respectively, using a gradient boosting classifier. Finally, this study presents the most important features acquired from the best performing classifier.

1 INTRODUCTION

The proliferation of fake publications in science is a significant concern, particularly in light of recent reports of large-scale retractions of publications from the permanent scientific record (Sabel et al., 2023). It is therefore necessary to develop effective methods for identifying unpublished manuscripts and publications that contain fraudulent content in order to maintain the integrity of science. In light of these developments, we sought to investigate the means by which fake publications can be identified. To this end, we searched for features of fake publications. Such articles that have been either withdrawn by the authors of the paper or by the editor of a journal or conference. There are numerous reasons why an article may be retracted. These include instances where a genuine mistake has been made by the authors, which has been identified only after publication and subsequently corrected. Alternatively, retraction may occur due to a fake peer review or as a result of manipulation of an image (Shen, 2020) or data (Oksvold, 2016). These kinds of papers are of significant concern in this research.

The objective of this study is to classify fake papers. But there are different aspects that label it a fake. Some features of a fake paper can include plagiarized images, paper written by a papermill, fabrication of results, and many more.

A faked paper can have either one or multiple reasons. However, there are some aspects of papers that are subjective from person to person; e.g., a paper written by a third-party writing service does not mean that the idea of experiments in the paper is fake, or the data is fabricated. Similarly, if a certain part of the paper has been written by AI or the results are not reproducible because the authors did not provide the data, this also does not mean that the whole study is fake. Therefore, in order to conduct this study, a proper definition of a fake paper needs to be established beforehand.

The papers used in the analysis of this study are deemed faked because they have been retracted for certain reasons discussed further in the study. Additionally, due to the limited availability of such fake papers, the reasons for deeming a paper fake have been selected from the list of reasons provided by Retrac-

[a] https://orcid.org/0009-0002-8223-4638
[b] https://orcid.org/0000-0002-4472-5543
[c] https://orcid.org/0000-0002-9484-1032
[d] https://orcid.org/0000-0003-4311-0624

Hussain, A. K., Sabel, B. A., Thiel, M. and Nürnberger, A.
Automated Detection of Fake Biomedical Papers: A Machine Learning Perspective.
DOI: 10.5220/0013482800003929

In *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025) - Volume 1*, pages 662-670
ISBN: 978-989-758-749-8; ISSN: 2184-4992

tion Watch [1]. The reasons selected are discussed in Chapter 2

Although the problem of fake papers is prevalent across various domains, its impact in the biomedical domain is particularly harmful for society, as a fake paper could have a profound effect on human health. Another reason to examine this area is that Retraction watch lists a significant majority (67%) of biomedical papers with at least one reason indicating that the paper is fake (Reasons discussed in Chapter 3). This illustrates the prevalence of the issue within this field and the necessity of addressing it. Moreover, fake papers are also cited in other publications. As described by (Bar-Ilan and Halevi, 2017), despite the clear retraction notices on the publisher's website, articles retracted in 2014 have been cited in 2015 and 2016.

The scale of the problem of increasing retractions in the biomedical domain is quite significant. To illustrate the scale of the problem (Gaudino et al., 2021) analyzed 5209 articles published between 1923 and 2020. Of the proportion of retractions, 83.8% were from clinical medicine, and 62.3% were due to scientific misconduct. Similarly, (Noorden, 2023) reported that Hindawi, a publisher of numerous medical journals, had retracted more than 10,000 papers in 2023 alone, an all-time high for the publisher.

The motivation behind this study is to address the problem using machine learning to develop and evaluate a series of classifiers that could automatically flag papers based on a specific feature set. The aim is to develop a classifier that can be used to reliably predict whether a paper could be a fake or not. Such a tool could prove beneficial to publishers, allowing them to subject such papers to a more rigorous scrutiny prior to their acceptance for publication.

This research has three main focus areas. The initial phase of the research involves the investigation of the performance of machine learning models in identifying retracted papers and comparing them with different evaluation metrics, including accuracy, precision, and recall. Secondly, it tests and validates different types of feature sets from publications, including metadata and content-based features and textual features, to identify which feature set performs best. Finally, it analyzes the key features used to make a prediction.

2 RELATED WORK

A review of the literature revealed research papers that address different aspects of fakeness of papers. For instance, (Razis et al., 2023) determines the probability of a paper being produced by a papermill or generated by AI by using a case-sensitive BERT-based model to classify the abstracts and titles of the papers with a recall score of 100% and a precision of 94%. The use of AI-generated text is a subject of debate, with ongoing discussions regarding the extent to which it should be permitted in academic literature. The question this raises is whether a paper that has been partially or completely written by ChatGPT should be considered a fake paper. In an article published in by Nature (Stokel-Walker, 2023) the ethical implications of four academic papers (Kung et al., 2022), (O'Connor and ChatGPT, 2023), (Transformer and Zhavoronkov, 2022), (Gpt and Steingrimsson, 2022) where language models were listed as authors of the papers are discussed. (Theocharopoulos et al., 2023) employs multiple different methods to detect AI-written abstracts, including LSTM+BERT, LR+TF-IDF and SVM+TF-IDF. However, the best results achieved were with LSTM+word2vec with a recall score of 98.6%. Similarly, (Desaire et al., 2023) conducted various experiments to detect ChatGPT-written text, by assigning it a role of a chemist. They achieve near 99% accuracy with their model for GPT 3.5 and 4 generated texts by using an XGBoost model for classification with 20 textual features, including the number of sentences per paragraph, presence of parentheses, and the presence of connecting words like 'although', 'because' etc. A drawback of this study is that the feature extraction process was conducted manually, and the analysis was limited to metadata features, not textual ones.

The inclusion of images that have been manipulated or tampered with in a research paper renders the paper as a whole as being of a lower standard of academic rigor. (Bucci, 2018) proposes a methodology to spot fake images in papers. The pipeline includes image extraction from the PDF, extraction of sub-image panels, multiple checks on images to look for duplication or manipulation. Elizabeth Bik (Shen, 2020) is a manual spotter of fake images, usually relating to western blot and microscopic images.

(Williams and Giles, 2015) has conducted a textual analysis to detect fake papers generated by SCIGen, a program that automatically generates scientific papers. They used different textual features, including key-phrase features, shingle features, simhash features, and TF-IDF features. The study reports a recall value of 0.999, for the TF-IDF features, but at the cost of a relatively low precision value of 0.251. Another study to detect fake papers generated by SciGEN (Xiong and Huang, 2009) that uses the references of the paper to verify the authenticity of the pa-

[1] retractiondatabase.org (Accessed at 5:50pm CEST on 16 September 2024)

per. It utilizes LAMP and Yahoo Boss OpenAPI to verify if the references actually exist or not.

Another characteristic of fake publications is miscitations, which refer to the use of a citation to assert a claim that is not explicitly mentioned in the referenced paper. (Liu et al., 2024) proposes a method for the detection of miscitations using a cosine similarity for the referenced paper and the context text where it was cited using sentence embeddings from BERT. The study reported an accuracy of 93% with a balanced dataset of 200 citations, using the complete abstract of the cited paper.

In addition to the aforementioned methods, various tools are available from companies for publishers to utilize in the screening of papers. Two notable examples are Integrity Hub by STM (International Association of Scientific, Technical and Medical Publishers (STM), nd) and the Papermill alarm (ClearSkies, nd). However, these tools are limited for publishers to use.

Additionally, manual methods have been proposed by (Sabel et al., 2023) and (Byr, 2020)for identifying indicators in manuscripts, including 'Chinese authors', 'Hospital affiliation', 'Requesting authors for full data'. Another study that employs manual methods for detection is (Dadkhah et al., 2023) which utilized a decision tree approach to classify fake papers. However, the feature extraction process was conducted manually, and the analysis was limited to metadata features, not textual ones.

The aforementioned studies address different aspects that constitute a fake paper. However, our objective is to provide a framework that scrutinizes a manuscript based on more than one indicator, thereby indicating whether a paper is suspicious. We aim to establish a strong baseline with relevant features from fake papers, using initially simple features to evaluate their performance. Furthermore, as observed from these studies, TF-IDF based features appear to perform well in comparison to other text-based features. Therefore, we explored the use of TF-IDF-based features as well as various BERT embeddings with different machine learning methods to assess how well fake papers can be classified. This was deemed of value because no existing studies have explored the feature sets and algorithms on fake papers. This study should therefore serve as a baseline for future research in the field of fake-publication detection in science.

3 DATASET COLLECTION AND CLEANING

The dataset was selected to include only papers from the biomedical domain, as the majority (67%) of retractions, due to reasons that make them fake, listed in Retraction Watch, originate from this domain. To this end, we created a new dataset with fake and non-fake papers in the biomedical domain. The fake paper DOIs were collected from Retraction Watch, a database of retractions from all scientific disciplines. However, we restricted our study to the biomedical domain by filtering the DOIs and choosing the categories (BLS) Biology and (HSC) Medicine. Secondly, as Retraction Watch also lists the reason for the retraction of a paper, we selected and analyzed those that indicated the possibility of them being fake. The list includes, but is not limited to, reasons such as 'Papermill', 'Concerns/Issues about Images', 'Plagiarism of Image', 'Duplication of data', 'Fake peer review' These reasons represent only a small subset of the 41 reasons (complete list provided in code) that were used to select the fake papers. It is worth noting that paper mills are agencies that fabricate fake scientific publications (Sabel et al., 2023).

These reasons for retraction were selected because we did not want to analyze retracted papers that had been retracted for valid reasons such as 'Not presented at conference' or 'Withdrawn to publish in different journal'. These criteria were deemed insufficient to indicate that these papers were fake. Subsequently, the papers were selected from the top 20 journals with the highest prevalence of fake papers. The rationale for selecting the top 20 journals was to eliminate the topic bias of different journals in selecting the control set and also selecting the journals that have been infested by fake papers. The fake papers were then filtered on the basis of the presence of a PubMed ID to avoid any potential bias, given that the non-fake papers were collected exclusively from PubMed. In order to select the non-fakes, it was necessary for them to be about similar topics to those covered by the fakes. Consequently, the non-fake papers were collected by searching PubMed for the keywords that had been derived from the titles of the fakes. The TfidfVectorizer from sklearn was employed to identify the top 150 important words from the titles of the fake papers. These keywords were used to search for non-fake papers. In order to further avoid bias, both classes of papers were selected from the period between 2012 and 2021, as there the number of fake papers published in these years was relatively higher than in other years. The final dataset consisted of 4634 fakes and 6624 non-fakes. The

metadata for all the papers and the titles and abstracts were extracted using Elsevier Scopus API (Elsevier, 2024) using the DOIs. The final dataset and code can be accessed in the linked repository.[2]

4 EXPERIMENTS

This section outlines the experimental design and methodology adopted to classify fake papers. The workflow for the classification is illustrated in Fig. 1. We conducted a number of classification experiments using various types of machine learning algorithms and features/indicators with the objective of gaining insights into the potential utility of these features for classification. Firstly, a feature set is needed to train a machine learning classifier. To this end, a number of different types of features and their combinations were evaluated to check which ones performed best. Secondly, we calculated the feature importances to analyze the most important features required to distinguish the classes. The following subsections present an overview of the methods used to extract the features and use them in various models for binary classification.

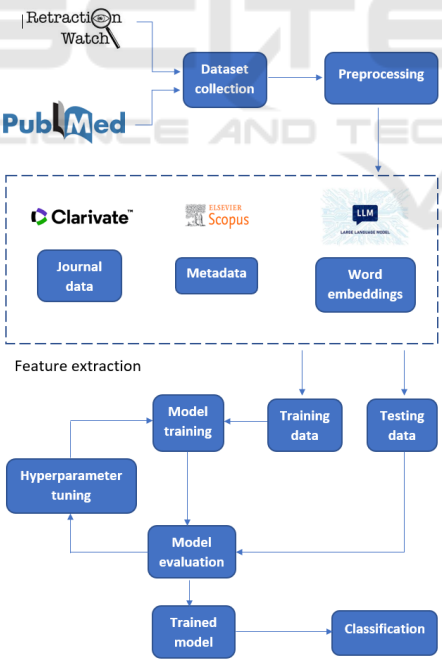


Figure 1: Workflow for classification.

[2]https://anonymous.4open.science/r/Classification-of-fake-papers-in-biomedicine-with-machine-learning-BBED/

4.1 Metadata and Journal Features

Firstly, the metadata features were explored in order to identify potential patterns that could differentiate fake papers from non-fake ones. The list of metadata features was extracted from publications using Scopus API shown in Table 1. The journal data acquired from Journal Citation Reports by Clarivate (cla, 2024) is presented in Table 2. The complete list of the journal features can be found in the code. It should be noted that some of the features in the Table 1 have been labelled as dummy features in order to avoid disclosing the features to papermills. The complete feature list can be acquired upon request from the authors.

Table 1: Metadata features.
Legend: N: Numerical, B: Binary.

Feature name	Description	Type
Dummy feature 1	Dummy feature	N
Number of authors	Number of authors of the paper	N
Open access	If the paper is open access	B
Dummy feature 2	Dummy feature	B
Hospital affiliation	If the authors have the words hospital associated with them	B
Country affiliation	The country affiliation of the first author	B
Title word count	Number of word in the title	N
Dummy feature 3	Dummy feature	N

Table 2: Journal features.

Feature name	Description
Journal	Name of the journal the paper was published in
Total citations of journal	Total citations from the lastest year JCR has data available for
Journal impact factor (JIF)	Average number of citations received by articles published in the last two years
% of articles in citable items	Percentage of items that can be cited
Journal immediacy index	Count of citations in the current year to the journal that reference content in this same year
Citable items	Items that contribute to the impact factor e.g. articles, reviews

The 'Hospital affiliation' feature is only relevant to biomedical papers and is a significant indicator of a fake paper, as previously reported by (Sabel et al., 2023). Consequently, we have also included this feature in our analysis. The rationale for including 'Number of authors' as a feature is to test the proposition that papers produced by papermills would not typically have fewer authors, given that a papermill would seek to sell the authorship of the paper to multiple authors to generate revenue and split the fee amongst them. The title word count is also included to ascertain whether fake papers authored by papermills exhibit a distinctive pattern, which would not be used by a non-fake paper. Finally, we use the open access indicator to check the prevalence of openly accessible retractions. The remaining features are related to the journal and were included to test the influence of the journal metrics on the classification task. Using the above mentioned metadata and journal features, a number of different machine learning classifiers were trained to classify fake and non-fake papers. The re-

sults of these experiments will be discussed further in chapter 5.

4.2 TF-IDF-Based Features

As discussed in the related work, (Williams and Giles, 2015) reported promising results for fake paper detection using TF-IDF (Salton and Buckley, 1988) based features. Accordingly, this avenue was explored to analyze whether textual-based features in fake papers exhibit a distinctive pattern of writing style or a repetitive use of vocabulary, which could be detected. Hence, we use TF-IDF scores from the abstracts of papers.

Prior to using the TF-IDF vectors, the abstracts+titles were pre-processed in order to remove irrelevant information. The first step of preprocessing involves the removal of stopwords. The stopwords that were removed were the standard English stopwords from the NLTK library (Bird et al., 2009). Subsequently, the data is tokenized using the word_tokenize functionality from NLTK. Finally, lemmatization is carried out as well in order to merge features that are essentially the same word but in a different form. This process ensures that a large, sparse feature set is not produced, with the vocabulary of the words used in the abstracts as features. Subsequently, the TF-IDF feature set is then employed to train a variety of classifiers.

4.3 Word Embeddings

Another popular approach of representing text for machine learning techniques is word embeddings. We use different models, including word2vec (Mikolov et al., 2013) and a number of BERT models, to generate sentence embeddings from the abstracts and the titles of the papers for classification. The BERT models used were pre-trained on medical texts, including, BioBERT, ClinicalBERT, PubMedBERT, SciBERT, BlueBERT, BioClinicalBERT from the transformers library from Hugging face (Face, 2023). The embeddings produced are sentence embeddings by averaging out the embeddings for individual words in the abstract+title. The rationale behind using sentence embeddings was to capture specific sentence structures and vocabulary that are frequently used by fake papers. The results of classification using sentence embeddings are shown in chapter 5.

4.4 Combining Features

This experiment includes the combination of different feature sets. We combine the metadata and journal features with the TF-IDF and the sentence embedding features to test if the performance of the classifiers could be enhanced. Although we only have a small number of metadata features compared to the high number of TF-IDF features or word embeddings, the classifiers that we selected for the study are able to evaluate feature importance and ignore the irrelevant features.

4.5 Classifiers

We used a number of different machine learning-based classifiers to test their performance for this problem, including Logistic regression (LR), Naive Bayes classifier (NB), Random Forest classifier (RF), Gradient boosting classifier (GB) and Decision trees (DT). Tree-based classifiers are of particular interest as they provide feature importance scores, which represent a significant aspect of our research. For all the classifiers, the data was split: 70% was used for training and 30% for testing. Moreover, the training set was split into 80% training and 20% validation set. Additionally, we conducted a 5-fold cross-validation on the training set. Each classifier was trained on a distinct subset of feature combinations. For each classifier, hyperparameter tuning was carried out using GridSearchCV(Pedregosa et al., 2011) and evaluated on the validation set. For all classifiers, the model was ultimately evaluated on the test set in order to compute the evaluation metrics.

4.6 Full Text Analysis

The final experiment in this study utilizes the full texts of the papers, i.e., from the introduction up to, but excluding, the references. The rationale for using features from full texts of the papers was to obtain further information about fake papers, i.e., some differentiating features that, alone with metadata or abstracts, cannot be found. We conducted this experiment with a shorter version of the dataset, as the majority of the papers in the original dataset were either not open source or could not be retrieved without the manual effort of downloading the full text. The smaller dataset consists of 1134 fakes and 3098 non-fakes. The Gradient boosting classifier was used in this experiment, as it showed promising results in the previous experiments. A subset list of features used is shown in Table 3. A full list of the features can be found in the code.

Table 3: Full text features.

Feature	Description
Flesch-Kincaid Grade Level	Readability score
Active Voice %	Percentage of active voice
Lexical Density	The proportion of content words to function words
Stopword ratio	Ratio of stopwords to total words
Hedging term frequency	Words like "might", "possibly", or "suggests"
Modal verb frequency	Modal verbs like 'can', 'will' and 'should'

The intuition for choosing readability score as a feature is that the fake papers might overly use unnecessary complex language to sound more 'scientific', thus increasing the Flesch-Kincaid score. Percentage of active voice is used as a feature because the fake papers might use less active voice compared to the genuine ones in order to obscure responsibility and make vague claims to avoid scrutiny. Lexical density is used as a feature because if a fake paper contains excessive filler text or redundant phrasing, its lexical density would be lower. The stopword count could be higher in fake papers because they would attempt to reach the word count, whereas, genuine papers would use precise terminology and minimize unnecessary words. The reason for using features such as hedging and modal verb frequency is because they acknowledge uncertainty in scientific text, therefore authors of fake papers might use them more than the non-fakes to avoid making definitive statements that could be easily challenged.

Along with these features, LLMs were also used to produce sentence embeddings from the full texts for the purpose of classification. However, most of the LLMs used have a context window of only 512 tokens. To address this limitation, two methods are used. The first method is to chunk the whole texts of the papers to fit in the context window and produce an average embedding of the chunks. The second method is to classify on the basis of the individual chunks and take a majority vote of all the chunks of a document. The results of this experiment are shown in Chapter 5

5 RESULTS AND DISCUSSION

The results presented in Table 4 for the models are tuned for the best-performing hyperparameters with recall score as the evaluation metric. The details of the hyperparameters can be seen in the code. Although there always remains some degree of uncertainty for a given manuscript about its authenticity, our method is valuable because the consequences of a potential fake classified as a non-fake would be more severe than a non-fake classified as a fake, which could be further scrutinized by human inspection.

A comparison of the results in Table 4 with the evaluation metrics for all feature sets reveals that a mixed feature set, which contains both metadata and journal-based along with textual features, performs best for all metrics across all algorithms. In our experiments, the Gradient boosting classifier performed the best with a recall of 83% and an accuracy of 86%, while also maintaining a high precision of 84%. In terms of all the metrics, the Gradient boosting classifier has outperformed the rest of them. The addition of the textual features to the metadata enhances the performance significantly, which may suggest that the fake papers have a typical vocabulary usage or sentence structure. In order to investigate further, the most important features for the top-performing classifiers are analyzed in the next section.

A number of different textual features were evaluated to check which one provides the best results. The results of the models used can be seen in Table 5. The results are in combination with the metadata features, and they show that BioBERT performs the best with a recall score of 83 %. BioBERT is a model trained on PubMed abstracts of papers.

5.1 Feature Importance Analysis

The sklearn implementation of tree-based machine learning models provides functionality to calculate the feature importances using the feature_importance_ method. Feature importances for best-performing tuned classifier: Gradient boosting classifier were calculated in order to investigate the features that were crucial for the separation of the classes. The feature importances are normalized so that the sum of all importances equals 1.

We selected the mixed feature set because it performed the best in terms of all evaluation metrics. The 10 top important features were analyzed and are shown in Figure 2.

The features presented in Figure 2 are arranged in descending order of their importance scores. The red bars illustrate the metadata and journal-related features, whereas the blue bars represent the TF-IDF features. It can be observed that among the top important metadata and journal features, having a China affiliation is the most important feature in making a classification, followed by the Dummy feature 1 of the paper. Among the TF-IDF features, the most frequently occurring term contributes the highest importance at around 3% for the classification.

To further visualize the differences among the top important metadata binary features in the fake and non-fake papers, Figure 3 presents a stacked bar plot of 'China affiliation' of the data across the two classes. The feature is plotted by normalizing both

Table 4: Evaluation metrics of algorithms across different feature sets.

	Metadata				BioBERT				Mixed			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
LR	0.72	0.68	0.62	0.65	0.78	0.76	0.70	0.73	0.71	0.67	0.60	0.63
NB	0.52	0.46	0.85	0.60	0.71	0.63	0.72	0.67	0.50	0.44	0.84	0.58
RF	0.81	0.80	0.71	0.76	0.78	0.77	0.66	0.71	0.84	0.83	0.76	0.79
GB	0.82	0.81	0.72	0.76	0.79	0.76	0.71	0.74	**0.86**	**0.84**	**0.83**	**0.83**
DT	0.75	0.70	0.70	0.70	0.65	0.58	0.60	0.59	0.76	0.69	0.74	0.71

Table 5: Evaluation metrics of algorithms across different embeddings.

	Acc.	Prec.	Rec.	F1
TF-IDF	0.86	0.84	0.81	0.82
Word2vec	0.85	0.83	0.81	0.82
BioBERT	0.86	0.84	0.83	0.83
ClinicalBERT	0.85	0.82	0.80	0.81
PubMedBERT	0.85	0.82	0.81	0.81
SciBERT	0.86	0.84	0.82	0.83
BlueBERT	0.85	0.83	0.82	0.82
BioClinicalBERT	0.85	0.82	0.80	0.81
BioGPT	0.86	0.83	0.83	0.83

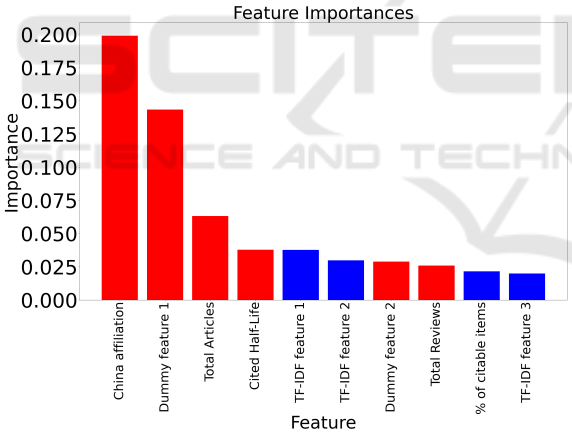


Figure 2: Bar plot of top 10 most important features for Gradient boosting classifier.

classes so that the proportion of the feature can be expressed as a percentage. The vertical axis in Figure 3 illustrates the proportion of the feature present in the data across the classes. It can be observed that around 75% of the fake papers have an affiliation with China, whereas among the non-fakes it is only 35%. The binary feature demonstrates a clear distinction between the output classes, demonstrating it plays a crucial role in distinguishing fake from non-fake papers in the classification process.

To demonstrate potential differences between the two classes in terms of the numerical metadata, box

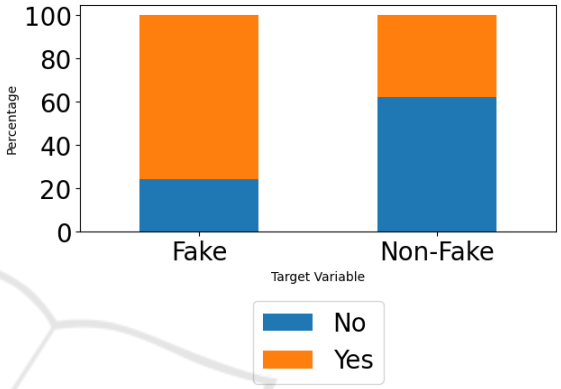


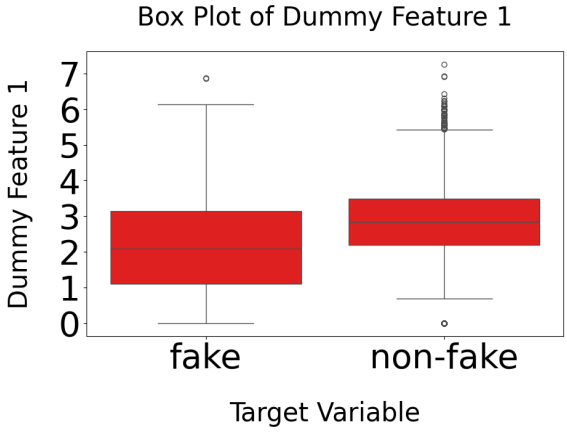
Figure 3: Percentage of papers with Chinese affiliation.

plots of the distribution of features were plotted in Figure 4. The vertical axis has been log transformed to reduce skewness and facilitate better visualization of the plots. For numerical variables such as 'Dummy feature 1' and 'Total Articles' a difference in the median values and variability across the classes can be observed. This suggests that these features are important and contribute to a certain extent in the distinction between the two classes.

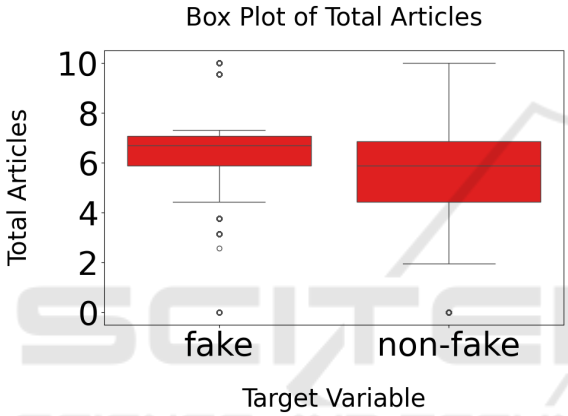
To visualize the TF-IDF features, as there are too many to plot, a word cloud was constructed that visualizes the top 74 most important features. In order to avoid providing information that could be used to educate papermills, we have not included the visualization of the TF-IDF features in this paper. However, we can send it to trustworthy fellow researchers upon request. The top features in the cloud indicate that the fake papers use these words commonly, and most of the fake papers are from specific fields in biomedicine. Other than that, the cloud contains verb usage as well, which indicates their common usage in fake papers.

5.2 Full Text Analysis

The results for the full text analysis from the Gradient boosting classifier can be seen in Table 6. The results



(a) Dummy feature 1.



(b) Total articles.

Figure 4: Log-transformed box plots for numerical metadata and journal features across classes.

demonstrate that the performance with features from full texts of the documents is relatively poor. The reason for this could be less data to train on or word embeddings with too much variability in the structures of the documents.

Table 6: Evaluation metrics for full text features.

	Accuracy	Precision	Recall	F1
Full text features	0.76	0.60	0.37	0.45
SciBERT with majority vote	0.77	0.63	0.33	0.44
SciBERT with summarization	0.74	0.46	0.38	0.41

6 CONCLUSION

Using machine learning, this study was designed to develop a classifier that can be used to flag papers as being fake and explore its features. We explored a number of different feature sets and classifiers, and finally investigated important features involved in separating fake papers from non-fake ones. We concluded that the combination of metadata and journal-related features with BioBERT embeddings provides better classification performance (83% recall) compared to the individual feature sets when taken alone. Secondly, the study concluded that the metadata feature 'Affiliation country as China' as well as certain biomedical vocabulary and verb usage prove to be strong indicators to flag a biomedical manuscript for further screening prior to publishing. It should be noted here that the specific domain vocabulary usage is dependent upon the popular research topic in that time frame. The classifier would need to be updated with time as research topics evolve.

While this study has advanced our understanding of how to identify detection of fake papers, there are still other avenues that are yet to be explored. This paper considers only a subset of features from the publication, and it would be valuable to explore additional features, such as sentiment analysis, tortured phrases, and image patterns. The textual features employed in our study were also limited to the abstracts of the papers due to the unavailability of most of the full texts. Lastly, our study was limited to publications in the biomedical field. To what extent the principles we uncovered are applicable to other fields or generalized to encompass all fields in science needs to be explored. In sum, using machine learning, it is possible to fake publications, and this knowledge might be useful for screening them.

ACKNOWLEDGMENT

This study is a part of the FASCIFFT I 167 (Fake science journals and their techniques) project at the Faculty of Informatik at Otto von Guericke University, Magdeburg, Germany. The project is funded by the state of Saxony Anhalt, Germany.

REFERENCES

(2020). Digital magic, or the dark arts of the 21st century—how can journals and peer reviewers detect manuscripts and publications from paper mills? *FEBS letters*, 594(4):583–589.

(2024). *Journal Citation Reports*. Clarivate Analytics. https://jcr.clarivate.com/, accessed September 20, 2024.

Bar-Ilan, J. and Halevi, G. (2017). Post retraction citations in context: a case study. *Scientometrics*, 113(1):547–565.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Bucci, E. M. (2018). Automatic detection of image manipulations in the biomedical literature. *Cell Death & Disease*, 9(400).

ClearSkies (n.d.). Papermill alarm. https://clear-skies.co.uk/. Accessed: 2024-09-27.

Dadkhah, M., Oermann, M. H., Hegedüs, M., Raman, R., and Dávid, L. D. (2023). Detection of fake papers in the era of artificial intelligence. *Diagnosis*, 10(4):390–397.

Desaire, H., Chua, A. E., Kim, M.-G., and Hua, D. (2023). Accurately detecting ai text when chatgpt is told to write like a chemist. *Cell Reports Physical Science*, 4(11):101672.

Elsevier (2024). Scopus api.

Face, H. (2023). Transformers: State-of-the-art machine learning for pytorch, tensorflow, and jax. https://huggingface.co/transformers.

Gaudino, M., Robinson, N. B., Audisio, K., Rahouma, M., Benedetto, U., Kurlansky, P., and Fremes, S. E. (2021). Trends and Characteristics of Retracted Articles in the Biomedical Literature, 1971 to 2020. *JAMA Internal Medicine*, 181(8):1118–1121.

Gpt, A. O. T. and Steingrimsson, S. (2022). Can gpt-3 write an academic paper on itself, with minimal human input? {hal-03701250}.

International Association of Scientific, Technical and Medical Publishers (STM) (n.d.). Stm integrity hub. https://www.stm-assoc.org/stm-integrity-hub/. Accessed: 2024-09-27.

Kung, T. H., Cheatham, M., , Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., and Tseng, V. (2022). Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *medRxiv*.

Liu, Q., Barhoumi, A., and Labbé, C. (2024). Miscitations in scientific papers: dataset and detection. Preprint.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Noorden, R. V. (2023). More than 10,000 research papers were retracted in 2023 — a new record. *Nature*, 624:479–481.

Oksvold, M. P. (2016). Incidence of data duplications in a randomly selected pool of life science publications. *Science and Engineering Ethics*, 22(2):487–496.

O'Connor, S. and ChatGPT (2023). Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse? *Nurse Education in Practice*, 66:103537.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Razis, G., Anagnostopoulos, K., Metaxas, O., Stefanidis, S.-D., Zhou, H., and Anagnostopoulos, I. (2023). Papermill detection in scientific content. pages 1–6.

Sabel, B. A., Knaack, E., Gigerenzer, G., and Bilc, M. (2023). Fake publications in biomedical science: Red-flagging method indicates mass production. *medRxiv*, pages 2023–05.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Shen, H. (2020). Meet this super-spotter of duplicated images in science papers. *Nature*, 581:132–136.

Stokel-Walker, C. (2023). Chatgpt listed as author on research papers: many scientists disapprove. *Nature*, 613:620–621.

Theocharopoulos, P. C., Anagnostou, P., Tsoukala, A., Georgakopoulos, S. V., Tasoulis, S. K., and Plagianakos, V. P. (2023). Detection of fake generated scientific abstracts. In *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*, volume 33, page 33–39. IEEE.

Transformer, C. and Zhavoronkov, A. (2022). Rapamycin in the context of pascal's wager: generative pre-trained transformer perspective. *Oncoscience*, 9:82–84.

Williams, K. and Giles, C. L. (2015). On the use of similarity search to detect fake scientific papers. In Amato, G., Connor, R., Falchi, F., and Gennaro, C., editors, *Similarity Search and Applications*, pages 332–338, Cham. Springer International Publishing.

Xiong, J. and Huang, T. (2009). An effective method to identify machine automatically generated paper. In *2009 Pacific-Asia Conference on Knowledge Engineering and Software Engineering*, pages 101–102.